LES SONDAGES : DÉLAISSÉS PAR LES STATISTICIENS ET MALMENÉS PAR LES POLITOLOGUES

Jeanne Fine - IUFM Midi-Pyrénées
Laboratoire de Statistique et Probabilités, Toulouse III

PLAN

1. Introduction, vocabulaire
2. Marge d'erreur de 3 % du sondage par quotas ?
3. Sondage aléatoire simple à probabilités égales ; marge d'erreur
   - Remarque 1 : échantillon aléatoire simple "représentatif" pour *n* assez grand
   - Remarque 2 : défaut de couverture, estimation biaisée
   - Remarque 3 : redressement de l'échantillon
   - Remarque 4 : les non-réponses
   - Remarque 5 : estimation d'une proportion dans une sous-population
4. Échantillon aléatoire stratifié (proportionnel puis optimal)
5. Méthode empirique des quotas
6. Les sondages, délaissés par les statisticiens et malmenés par les politologues

1. Introduction, vocabulaire

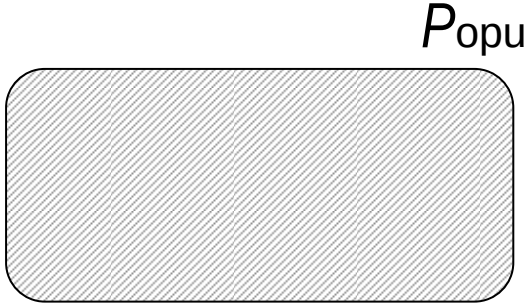
Publication dans *Le Monde* du 26 avril 2002 de deux articles :

- "Sondages et regrets" par Roland Cayrol, Directeur associé de l'Institut CSA
- "Faute de contrôles" par Michel Lejeune, Statisticien.

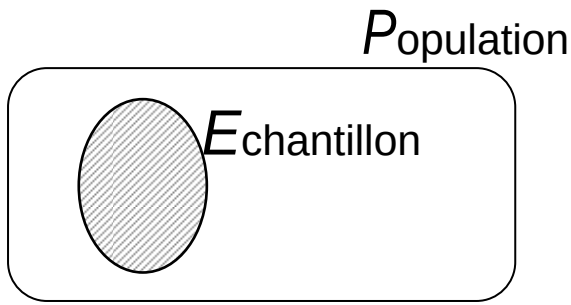
Vocabulaire :

- "enquête par sondage" et "sondage d'opinion"
- "hasard" et "aléatoire"
- "sondage" et "recensement"

Recensement
Toute la population est enquêtée



Sondage
Une partie de la population, l'*échantillon*, est enquêtée.



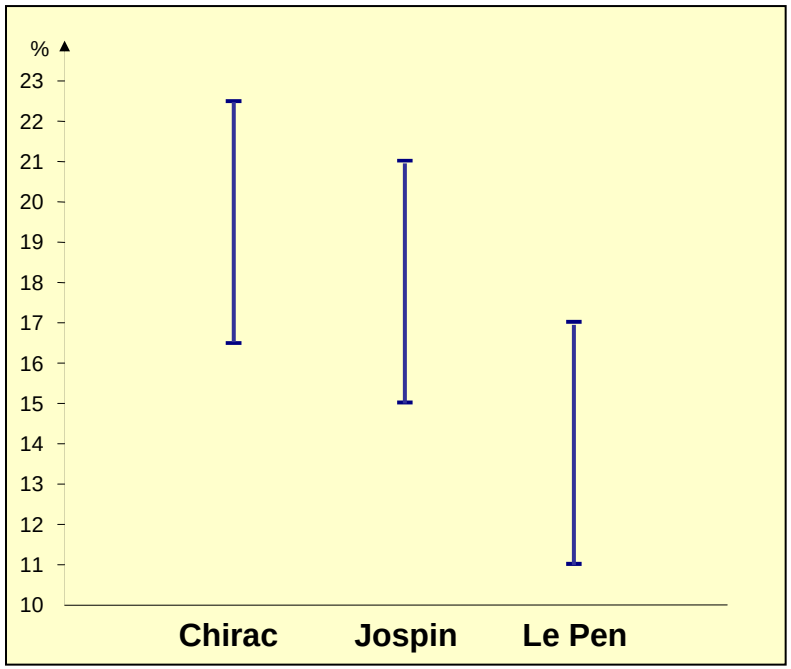
2. Marge d'erreur de 3% du sondage par quotas ?

Roland Cayrol : les sondeurs ne se sont pas trompés

	Sondage	Élection	Écart
Chirac	19.5 %	19.7 %	0.2 %
Le Pen	14.0 %	16.9 %	2.9 %
Jospin	18.0 %	16.1 %	1.9 %

Et si les résultats étaient annoncés avec une marge d'erreur de 3 % ?

Chirac	entre 16.5 % et 22.5 %
Jospin	entre 15.0 % et 21.0 %
Le Pen	entre 11.0 % et 17.0 %

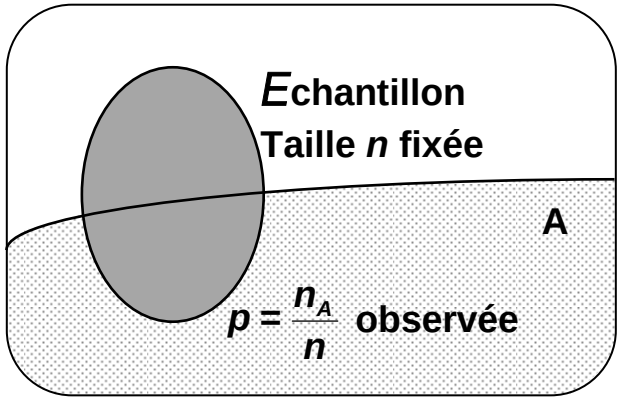


3. Sondage aléatoire simple à probabilités égales

Michel Lejeune :

Les derniers sondages indiquaient 18% pour Jospin et 14% pour Le Pen. Pour les rares scientifiques qui savent comment sont produites les estimations, l'écart rendait tout à fait plausible le scénario qui s'est réalisé. Si l'on se réfère à un sondage qui serait effectué dans des conditions idéales (tirage aléatoire absolu, taux de réponse 100%, aucune fausse déclaration) on obtient sur de tels pourcentages une incertitude de plus ou moins 3% étant donné la taille de l'échantillon.

3. Théorème



Théorème :

Soit *A* une partie de la population et *P* la proportion de *A* dans la population.
Soit *p* la proportion de *A* observée sur un échantillon aléatoire simple à probabilités égales de taille *n* (on suppose *n / N* négligeable*).*

Estimation par intervalle à 95 % de confiance de la proportion *P :*

$$\left[p - 2\sqrt{\frac{p(1-p)}{n}} ; p + 2\sqrt{\frac{p(1-p)}{n}} \right]$$

3. Marge d'erreur

Marge d'erreur à 95 % de confiance : $2\sqrt{\frac{p(1-p)}{n}}$

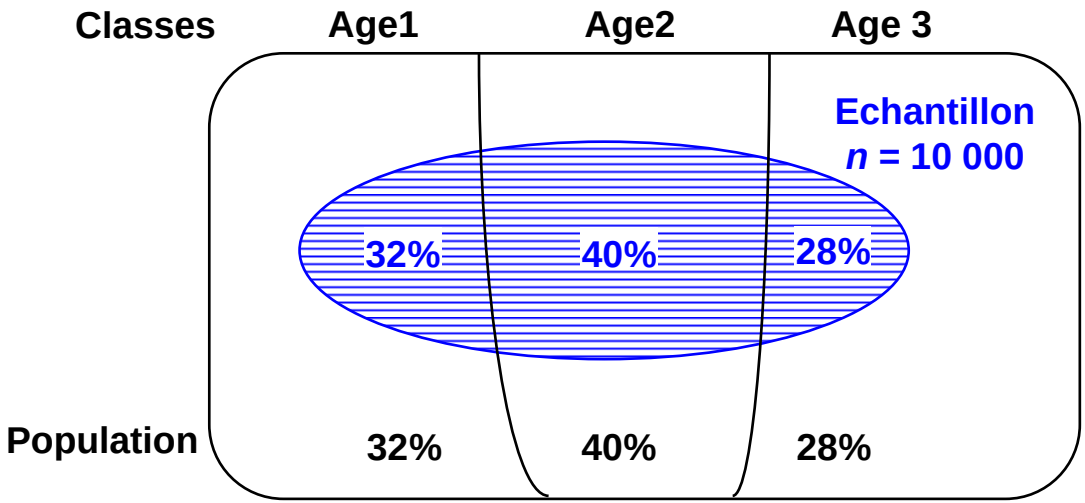
Proportion observée p Taille échantillon n	10% ou 90%	20% ou 80%	30% ou 70%	40% ou 60%	50 %
100	6.0 %	8.0 %	9.2 %	9.8 %	**10.0 %**
400	3.0 %	4.0 %	4.6 %	4.9 %	**5.0 %**
1 000	1.9 %	2.5 %	2.9 %	3.1 %	**3.2 %**
5 000	0.8 %	1.1 %	1.3 %	1.4 %	**1.4 %**
10 000	0.6 %	0.8 %	0.9 %	1.0 %	**1.0 %**

Pour $p \cong 50$ % l'intervalle s'écrit : $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ c'est-à-dire $p \pm \frac{1}{\sqrt{n}}$

Pour $n = 1000$, on a : $p \pm 3\,\%$

La fourchette ne dépend pas du taux de sondage n / N.

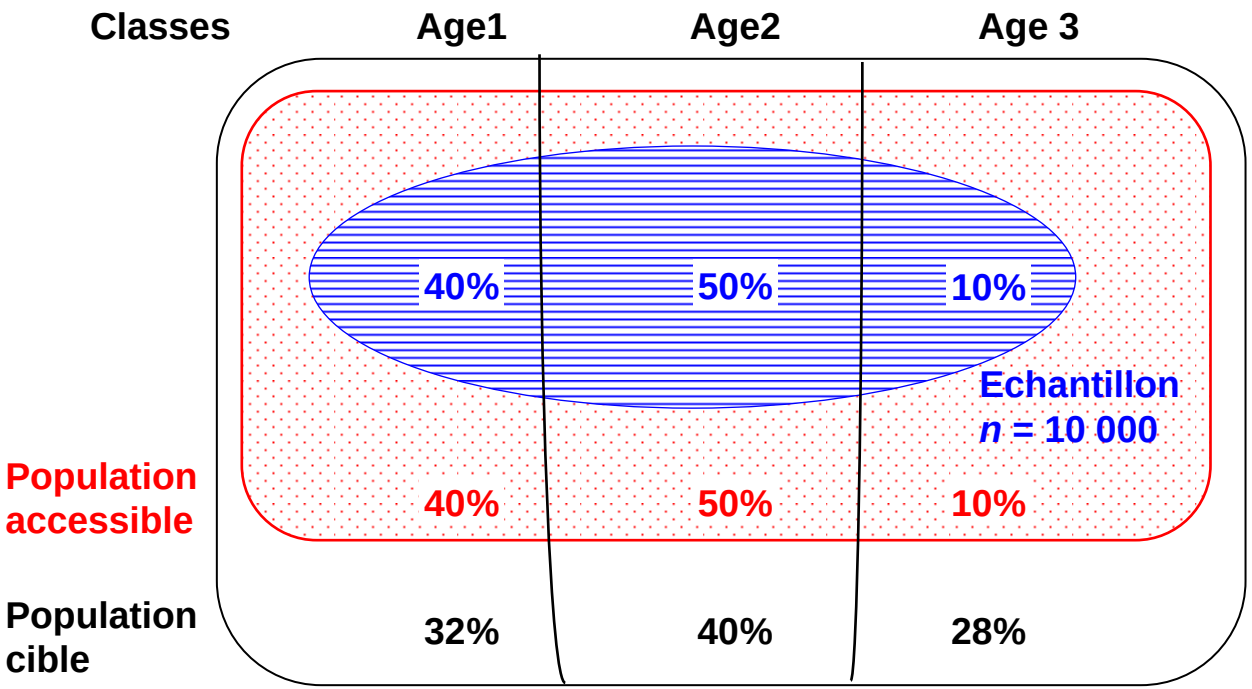
Remarque 1. Echantillon aléatoire simple "représentatif" pour *n* assez grand



Un échantillon aléatoire simple à probabilités égales de taille 10 000 est « représentatif » par rapport à toutes les variables connues ou inconnues sur la population.

Remarque 2. Défaut de couverture, estimation biaisée

Si la population *accessible* par les enquêteurs est différente de la population *cible* (*défaut de couverture*), on peut obtenir des *estimations biaisées*.



Pourcentage de voix pour X	
Pop. cible	52%
Pop.access. et échantillon	46%

Remarque 3. Redressement de l'échantillon

Répartition par âge

	Age 1	Age 2	Age 3	Ens.
Pop. cible	32%	40%	28%	100%
Pop. access. et échantillon	40%	50%	10%	100%

Pourcentage de votants pour X

	Age 1	Age 2	Age 3	Ens.
Pop. cible	50%	51%	56%	52%
Pop. access. et échantillon	45%	46%	54%	46%

$$0.45 \times 0.40 + 0.46 \times 0.50 + 0.54 \times 0.10 \simeq 0.46$$

Le *redressement d'échantillon* consiste à utiliser la répartition de la population cible au lieu de la répartition de l'échantillon ; on obtient :

$$0.45 \times 0.32 + 0.46 \times 0.40 + 0.54 \times 0.28 \simeq 0.48$$

Résultat encore biaisé : 48% est différent de 52%.

Remarque 4. Les non-réponses

On demande aux 600 élèves d'un lycée (300 filles et 300 garçons) s'ils ont déjà consommé de la drogue.

Résultats de l'enquête sur les 600 élèves

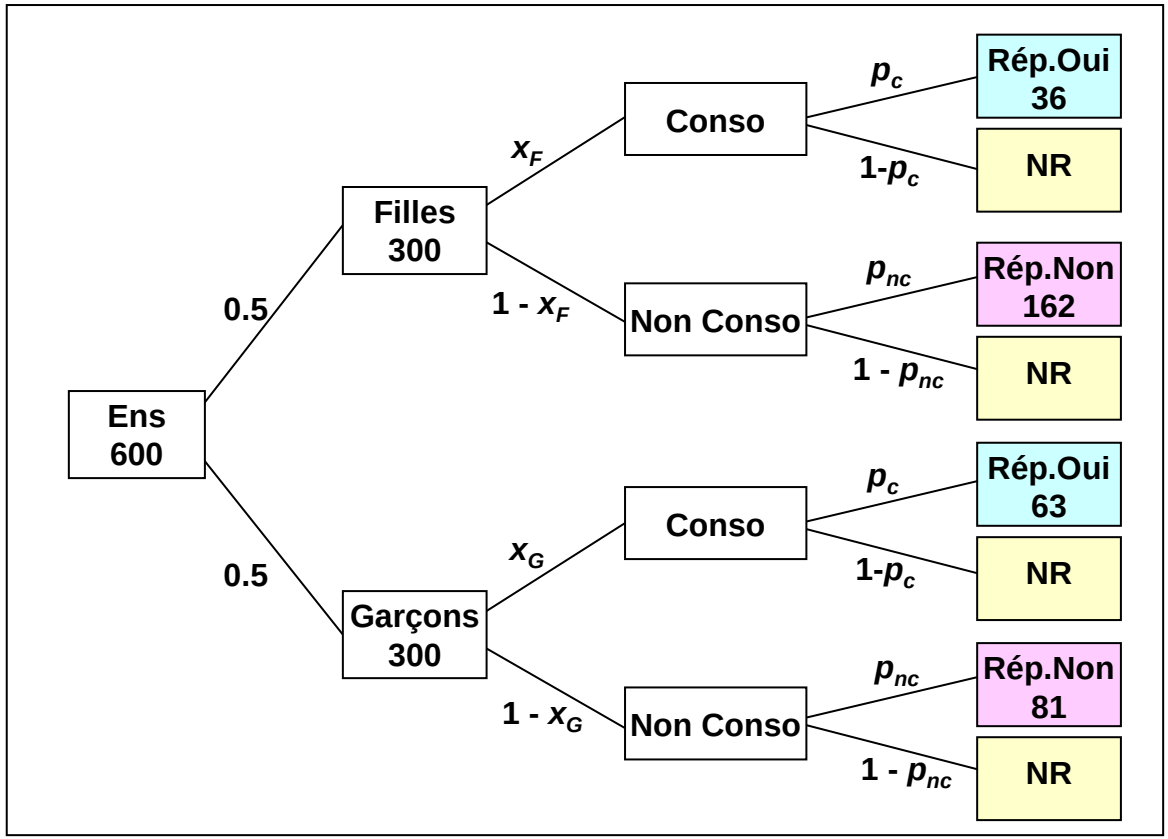
Consommation / Sexe	oui	non	Réponse	% de oui	Non Réponse	Ensemble
Filles	36	162	198	18%	102	300
Garçons	63	81	144	44%	156	300
Ensemble	99	243	342	29%	258	600

Notation et hypothèse :

x_F (resp. x_G) : proportion de filles (resp. de garçons) ayant consommé

p_c (resp. p_{nc}) : proportion des répondants parmi ceux qui ont consommé (resp. qui n'ont pas consommé)

Remarque 4. Les non-réponses



Des calculs simples permettent d'obtenir :

$$p_c = 0.3, \quad p_{nc} = 0.9, \quad x_F = 0.4 \quad \text{et} \quad x_G = 0.7$$

55 % des élèves du lycée ont déjà consommé de la drogue et non pas 29 % comme annoncé au départ en ne tenant compte que de ceux qui ont répondu.

Remarque 5. Estimation d'une proportion dans une sous-population

Enquête par sondage aléatoire simple à probabilités égales avec remise de taille 1000 :

15 % sont partis en vacances ;
parmi eux, 52 % sont partis à l'étranger.

Estimation par intervalle à 95 % de confiance :

$$15\% \pm 2\% \quad \text{mais} \quad 52\% \pm 8\%$$

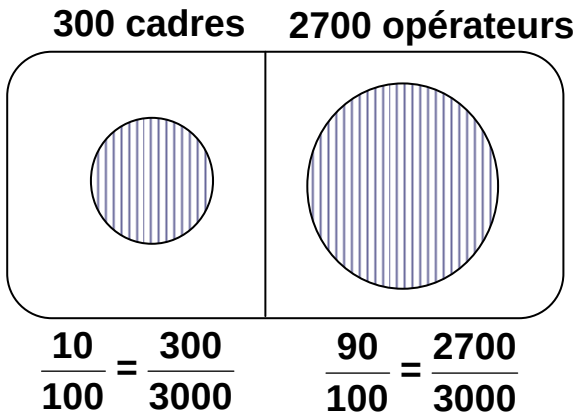
En effet, on a $2\sqrt{\frac{0.15 \times 0.85}{1000}} \simeq 2\,\%$ et $2\sqrt{\frac{0.52 \times 0.48}{150}} \simeq \frac{1}{\sqrt{150}} \simeq 8\,\%$

4. Échantillon aléatoire stratifié proportionnel

Estimation du salaire moyen d'une entreprise.
Echantillon aléatoire simple de taille 100 versus échantillon aléatoire stratifié *proportionnel* de taille globale 100 :

$$n_k \propto N_k$$

Échantillon stratifié proportionnel



À taille d'échantillon égale, l'échantillon aléatoire stratifié proportionnel donne des estimations plus précises que l'échantillon aléatoire simple ; d'autant plus que les moyennes par strates sont différentes.

4. Échantillon aléatoire stratifié optimal

À taille d'échantillon global fixée, la répartition de l'échantillon entre opérateurs et cadres qui donne la meilleure précision n'est pas l'échantillon proportionnel mais un échantillon qui tient compte également de la dispersion selon les strates de la variable objet de l'enquête :

$$n_k \propto N_k \sigma_k$$

L'amélioration de la répartition optimale sur la répartition proportionnelle est d'autant plus forte que les écart-types par strates sont différents.

L'échantillon stratifié est un échantillon à deux degrés, recensement au premier degré, les strates, sondage au deuxième degré.
On peut utiliser plusieurs degrés.

5. Méthode empirique des quotas

Michel Lejeune :

En pratique, on est loin de se trouver dans ces conditions : on ne peut pas réaliser un véritable tirage aléatoire parmi les électeurs, le taux de réponse est de 10 à 20 % en pratique dans le cas du téléphone, enfin, les fausses déclarations d'intention ne sont pas négligeables, en particulier sur les intentions de vote pour l'extrême droite. Les sondeurs croient, ou feignent de croire, que grâce à l'utilisation de quotas et au redressement des échantillons, leur précision pourrait être meilleure que, par exemple, ces 3 %.

5. Méthode empirique des quotas

		POPULATION	ÉCHANTILLON
SEXE	Hommes	48 %	480
	Femmes	52 %	520
	Ensemble	*100 %*	*1000*
ÂGE	< 35	27 %	270
	35 à 65	51 %	510
	>65	22 %	220
	Ensemble	*100 %*	*1000*
PCS	PCS +	33 %	330
	PCS –	30 %	300
	Inactif	37 %	370
	Ensemble	*100 %*	*1000*

Le grand avantage de la méthode des quotas est qu'elle ne nécessite pas de disposer d'une base de sondage d'où, comparativement à un sondage aléatoire de même taille, un très faible coût et une très grande rapidité. L'inconvénient est qu'il n'est pas possible de calculer la précision des estimateurs obtenus.

5. Méthode empirique des quotas

Roland Cayrol, dans son livre intitulé "Sondages, mode d'emploi", page 38 :

C'est le fameux exemple des boules rouges et des boules noires, tirées dans un sac qui comporte le même nombre de boules rouges que de boules noires.
Si l'on tire un échantillon de 100 boules du sac – mais attention, d'une manière purement et absolument aléatoire –, la théorie des probabilités indique qu'il y a 95 chances sur 100 (attention encore : pas cent pour cent) pour que le nombre de boules rouges ainsi tirées au hasard soit compris entre 45 et 55. Nous savons qu'il y a 50% de boules rouges dans le sac ; donc la proportion exacte de boules rouges aura été approchée, par le tirage d'un échantillon aléatoire, avec une marge d'erreur, un intervalle de confiance, de 10 % (45 à 55). Et il reste encore cinq chances sur cent que le résultat du tirage soit extérieur à cette fourchette.

Il est vrai que la marge d'erreur à 95% est de 10%,
ce qui donne un intervalle de confiance de la forme $50\% \pm 10\%$
et non $50\% \pm 5\%$ comme l'écrit Roland Cayrol !

5. Méthode empirique des quotas

Michel Lejeune :

Un sondeur a récemment déclaré que le vote Le Pen constaté dans les échantillons devait être multiplié (c'est-à-dire, redressé) par 2, … selon des règles mathématiques rigoureuses, ajoutait-il. En fait les corrections relèvent plus de l'appréciation empirique des politologues que de la théorie statistique. J'en veux pour preuve la surprenante proximité des résultats d'un institut à l'autre. Les sept derniers sondages publiés donnaient tous Jospin à 18% : pour tout observateur avisé, cette constance est statistiquement invraisemblable avec des échantillons de taille 1000.

On montre que la probabilité que sept sondages de taille 1 000 réalisés indépendamment et dans les mêmes conditions donnent le même résultat 18% à 0.5% près est égale à 3 pour 10 000 !

Il faut donc comprendre que ces résultats sont "moyennés et redressés" pour tenir compte de tous les sondages réalisés les derniers jours, ce qui explique que sondeurs et journalistes les commentent comme s'ils étaient quasi-certains (comme un échantillon aléatoire simple de taille 7 000).

5. Méthode empirique des quotas

Michel Lejeune :

Quant aux médias, ils sont les dupes, peut-être les complices, des discours lénifiants des instituts.

Ce sont principalement les médias qui commandent les sondages préélectoraux et qui les commentent.
Avant de converger de façon suspecte, les médias publient pendant des mois des sondages aux résultats contradictoires pourtant réalisés les mêmes jours.
Il n'y a qu'en France que sont publiés autant de sondages reposant sur une méthodologie aussi peu fiable.

6. Les sondages, délaissés par les statisticiens et malmenés par les politologues

Michel Lejeune :

Il faut savoir qu'il n'y a aucun statisticien spécialiste de la théorie des sondages dans les cinq instituts concernés. Aujourd'hui la communauté scientifique doit se sentir en partie responsable de ce qui vient de se produire. Par dédain, elle n'a jamais voulu s'intéresser à la pratique des sondages, ni vraiment d'ailleurs, à la théorie.